Mathematical Programming with Stochastic Equilibrium Constraints applied to Optimal Last-mile Delivery Services

B. Tounsi

University of Lille, INRIA, 40 Avenue Halley, Villeneuve d'Ascq, France

Y. Hayel, D. Quadri, L. Brotcorne

University of Avignon & NYU-poly, University Paris XI LRI, INRIA Avignon & New York City, Paris-Saclay, Villeneuve d'Ascq, France & USA

Abstract

In e-commerce business, the delivery of products is a crucial part for the success of an e-shop. An efficient delivery system should offer various services and predict customers behaviour. The latter are influenced by the price of a delivery service, but also by its quality (perceived through congestion effect induced by customers' choices). In this study, we introduce a bi-level model to optimize a delivery system. At the upper level, the provider control services' tariffs. At the lower level, users react by choosing their delivery service according to a disutility function which incorporates the provider tariff and the congestion effects. We model the customers' reaction using stochastic user equilibrium (SUE). We also present a sensitivity analysis for the SUE that gives explicit expression of the derivatives of customers distribution with respect to services' tariffs. Based on a local search that exploits the derivatives information, a new heuristic algorithm for a delivery services pricing problem is developed and compared to others existing methods.

Keywords: E-commerce, stochastic user equilibrium, bi-level programming, heuristics.

[1] This work is supported by the French National Research Agency (ANR-Agence Nationale de la Recherche) and is part of project RESPET ANR TTD 2011.

1 Introduction

In many contexts, optimizing a strategic system in which decision makers take decision selfishly on their own, is a challenging problem. This type of problem appears in many contexts like transportation networks, communication networks, economy, etc.. In this paper, we present a mathematical programming approach for solving a bi-level program with stochastic equilibrium constraint for a services' design problem. Our study is motivated by a problem of last-mile delivery services design for on-line retail [9]. After making their purchase, customers choose a delivery service offered by the provider. Each customer decides selfishly the best choice for him. Customers interact all together through congestion effects that are modeled considering queueing metrics. This type of interactive systems with large number of players competing into queueing systems is well overviewed in [8]. We are faced to a stochastic user equilibrium (SUE) problem that defines the lower level of our bi-level problem. At the upper level, the provider has to solve the services' design problem taking into account the equilibrium constraint. This problem appears difficult analytically as the cost functions are not simple and also there is no closed form expression of the dependence between the variables of the provider and the resulting customers' equilibrium. In the following we describe the delivery services and the customer' choice process in section 2. For determining the customer's equilibrium, we consider the method of successive averages (MSA) [14]. In section 3, the services' design problem is addressed and a sensitivity analysis for the SUE is given. We propose an efficient heuristic that is compared to two literature heuristics in section 4, along with suggestions of future works.

2 Delivery services and customers' choice

We assume that each customer chooses one service among a set $\mathcal{J}$ of services. The total number of services proposed is $|\mathcal{J}| = J$. We consider two families or types of services: delivery at home (DH) and delivery at warehouse (WH). The disutility (cost) of service j perceived by a customer depends on the tariff A_j set by the provider, and on the quality of service (evaluated on the congestion effect) induced by other customers' choices. We denote by λ the global customers arrival rate per unit of time. Particularly, we assume that this arrival process follows a Poisson process which means that the expected number of customers, at each time unit, is equal to λ. Each arriving customer makes a decision about his service and we denote by p_j the fraction of customers choosing service $j \in \mathcal{J}$. Note that we have $\sum_{j\in\mathcal{J}} p_j = 1$. The general form of

the disutility c_j of service j is given by the following function

$$c_j(p_j) = A_j + \alpha_j f_a(p_j)$$

where $f_a(p_j)$ is the congestion level of service j of type $a \in \{DH, WH\}$ and α_j a monetary conversion coefficient. The congestion function depends on the service. Indeed, the congestion is perceived in terms of the average delivery delay for the (DH) services, while it is related to the rejection probability for the WH services. In order to express the average delivery delay of a parcel $f_D(p_j)$ for a (DH) service j, we consider a general queue model which is an M/G/1. In fact, the arrival of packet is a Poisson process with rate λp_j as we are considering an unobservable strategic decision for each customer. Packets are delivered one by one, and the service time of each packet is a positive random variable. The average journey time (which corresponds to the average delivery delay) is given by the Pollaczek-Khinchin formula [9]. The congestion function $f_W(p_i)$ for a customer that chooses the (WH) service i is related to the fact that the parcel can be refused by the warehouse owner du capacity limit. This type of problem is understandable because most of warehouses are located in city center and their capacity storage is very limited. Moreover, we assume that a not rejected incoming parcel, stays a random amount of time in the warehouse before the destined customer comes to pick it up. This random time follows an exponential distribution with rate μ. Then, we model the (WH) service i as an $M/M/K_i/K_i$ queue where K_i is the capacity of the warehouse i. The congestion function $f_W(p_i)$ is expressed by the blocking probability and is given by the Erlang-B formula [9]. Once we have defined the disutilities for (DH) and (WH) services, we know look for a stable situation in which no customer has an interest to change his decision. It is more accurate to consider that users can make errors in there decision (lack of information, individual error,...). This type of assumption is widely proposed in Traffic Assignment Problems [14] and discrete choice theory in general [3]. It is known that it leads to the solution concept of Stochastic User Equilibrium (SUE) [1] and more generally to quantal-response equilibrium (QRE) [11]. In the latter general model, a QRE is a distribution vector $p^* = (p_1^*, \ldots, p_J^*)$ which verifies the following fixed-point system:

$$(1) \quad \forall j \in \mathcal{J}, \quad p_j^* = F_j(p^*),$$

where the function F_j is called the response function for a customers that chooses service j. Depending on the response functions considered, the QRE corresponds to well-known equilibrium concepts like Wardrop equilibrium, Logit equilibrium, Probit equilibrium, etc. The partial rationality assumption may come from an error ζ on the disutility that induced a bias in the

decision of each customer. In most DCM, the error is a random variable that follows one of the two distributions: Gaussian or Gumbel. The latter one is usually consider in TAP problems because it has many interesting properties like the *efficiency principle* [6]. Moreover, in our setting, we consider that services of the same type are correlated in the way that their probability to be chosen by any customer are not independent. In order to model this dependency and correlations between the services, we consider a nested Logit discrete choice model [3]. Services having the same type are merged into a nest, and the choice process can be seen as a two steps process for each customer. First, each customer determines the nest DH or WH; second, the customer choose one service of the nest. Using Bayes conditional probability, the choice distribution of any customer depend on the disutilities of services and parameters of the nested Logit model.
In the following proposition, we prove that a solution of this system can be obtained by solving a minimization problem. A similar result has been obtained for the case of the Logit SUE [7] and also recently in [13]. We formulate the proposition for a general nested Logit DCM with J choices and where each choice j is associated with a nest $n(j)$, with $\mathcal{N}$ the set of N nests. We denote by θ a positive coefficient related to the dispersion in the perceived disutilities, and by ϕ_n the correlation coefficient of nest $n \in \mathcal{N}$.

Proposition 2.1 *The Karush-Kuhn-Tucker conditions of the following minimization problem:*

(2) $min_p Z = Z_1 + Z_2 + Z_3$,
with:

$$Z_1 = \sum_{n\in\mathcal{N}} \sum_{j\in n} \int_0^{p_j} c_j(s)ds, \quad Z_2 = \frac{1}{\theta} \sum_{n\in\mathcal{N}} \sum_{j\in n} p_j \ln(p_j), \quad Z_3 = \sum_{n\in\mathcal{N}} \frac{1-\phi_n}{\theta\phi_n} \left((\sum_{j\in n} p_j) \ln(\sum_{j\in n} p_j) \right),$$

subject to the constraints:

$$\sum_{n\in\mathcal{N}} \sum_{j\in n} p_j = 1 \quad and \quad p_j \geq 0, \quad \forall j \in \mathcal{J},$$

are the conditions for a solution of the stochastic user equilibrium (1).

The existence of the solution for this minimization problem is proved by showing the convexity of the objective function Z. A similar equivalent minimization problem result has also been obtained for the cross-nested Logit model in [2]. Based on this proposition, we can apply well-known algorithms that have been proposed in the literature to solve this convex minimization problem.

Our setting is a particular type of TAP in which we can represent customers' choice process by a network with a parallel topology link and where each link is associated to a service. We note that the number of variables in our problem is limited compared to a general TAP. Thus we use a simple form of the method of successive averages (MSA) [14] to solve the SUE problem. The idea of this method is to start with an initial probabilities distribution and iteratively update services' disutilities and compute the new SUE. The algorithm stops when no change occur on the choices' probabilities. More efficient methods and technics are proposed recently in the literature like improving the MSA by optimal step sizes and directions [5], gradient/fast-gradient based method [13] and internal approches [4] to cite a few.

3 The services design problem

Having that customers follow a SUE, the provider is willing to optimal services' design regarding some criteria. This yields to a bi-level problem where several decision makers interact. More precisely, our problem is a *Mathematical Program with an Equilibrium Constraint (MPEC)* [10] where the equilibrium considered is stochastic. We assume in our framework that the provider, called the leader in a bi-level program, controls service tariffs A_j for all service $j \in \mathcal{J}$. For a particular set of the provider's variables, customers, the followers, react by choosing services in a way that minimizes the own perceived disutility. Denoting by u the provider's variables, the MPEC problem can be formulated as follows:

$$(MPEC) \begin{cases} max_u \ F(u,p), & (1) \\ \quad s.t \ L_j \leq u_j \leq U_j \ \forall j \in \mathcal{J}, & (2) \\ \quad p = p^*(u). & (3) \end{cases}$$

Where constraints (2) set bounds on provider's variables, and constraint (3) describes that p is a solution of the stochastic equilibrium problem parameterized by u. Similar MPEC formulation for traffic control application are addressed in [15] where a heuristic resolution method is proposed, while simulation is studied in [16] for a pricing problem under SUE. The main challenge for this optimization problem is to know how the stochastic equilibrium $p^*(u)$ depends on the parameter u. We present in the next section how we use sensitivity analysis to overcome this difficulty.

3.1 Sensitivity analysis for the SUE

As shown in previous section, the equilibrium $p^\star(u)$ is the solution of a nonlinear problem parameterized by u. For such problem, sensitivity analysis is illustrated in [15] as a tool to compute the variation of the solution when the

parameter is perturbed. This is explicitly done by computing an estimation of the gradient of $p^\star(u)$ with respect to u. In traffic assignment problem, sensitivity analysis for user equilibrium is carried out in the deterministic case [15] and for the Logit based SUE [12]. Following this idea, we derive the gradient of our nested Logit SUE with respect to the provider's variables. Precisely, for a current point $u^n = (.., u_i^n, ..)$, we compute the derivatives $[\frac{\partial p_j}{\partial u_i}]_{u=u^n}$, $\forall i, j = 1, .., J$.

We now introduce three heuristic methods for the services' design problem. While two of them (GDA and SLS) use sensitivity analysis, a third one (BLS) is implemented for benchmarking.

3.2 Gradient descent algorithm (GDA)

In this first approach, sensitivity analysis is used to compute the gradient of the objective function $F(u, p)$ by applying the following chain rule :

$$\frac{\partial F(u, p(u))}{\partial u_j} = \frac{\partial F}{\partial u_j} + \sum_{i \in \mathcal{J}} \frac{\partial F}{\partial p_i} \frac{\partial p_i}{\partial u_j}, \forall j \in \mathcal{J}.$$

The resulting gradient is used as descent direction to update services' tariffs. The detailed steps of GDA are as follows:

- Step 0. *Initialisation*: Init provider's variables u^0, and number of iterations $n = 0$.
- Step 1. Solve SUE problem parameterized with u^n using MSA.
- Step 2. Compute the derivatives $\frac{\partial p^\star}{\partial u}$, using the sensitivity analysis.
- Step 3. *Direction finding*: Compute descent direction d^n (gradient of F).
- Step 4. *Update*: $u^{n+1} = u^n + \sigma^n d^n$, σ^n is the step size.
- Step 5. *Convergence*: If $|u^{n+1} - u^n| < \epsilon$ then Stop, else go to step 1.

In step 4, we use a predefined step size $\sigma^n = 1/n$, while ϵ is some small value used as convergence threshold.

3.3 Bi-level local search (BLS)

This algorithm performs a local search starting from an initial set u_0 of provider's variables. In a neighbourhood V_0 of u_0, the objective function $F(u, p(u))$ is evaluated $\forall u \in \mathcal{V}_0$. The best neighbour $\bar{u}$ is picked and if it improves $F(u, p(u))$ it becomes the current point. New search is then started in the neighbourhood of $\bar{u}$. These steps are repeated until no improvement can be obtained. In the experimentation, at iteration n, a neighbour is build by adding or subtracting $\sigma^n = 1/n$ to each provider's variable. The major inconvenient of this heuristic is that at each neighbour of current point a SUE is computed using the MSA. To reduce computation time, we combine

in the next algorithm, local search technique with useful informations of the sensitivity analysis.

3.4 Sensitivity analysis based local search (SLS)

SLS is also a local search heuristic using the same neighbourhood than (BLS). The difference is that in the evaluation step, instead of computing a SUE for each neighbour, based on sensitivity analysis of section 3.1, we use the following approximation: $p_j^\star(u^{n+1}) = p_j^\star(u^n) + \sum_{i \in \mathcal{J}} \frac{\partial p^\star}{\partial u_i}(u_i^{n+1} - u_i^n)$.

4 Preliminary experimentations

We compare the three proposed algorithms with two examples and where the provider maximizes its revenue $F(u,p) = \lambda \sum_{j \in \mathcal{J}} u_j p_j$. In the first example, there are 3 services: one is a (DH), the two others are (DW) and they form a nest. One option of the latter is not controlled by the provider to avoid trivial solution. In the second example, the number of services is 6 distributed in two nests and enables us to well measure the performance of each algorithm.

	Ex 1: 3 services			Ex 2: 6 services		
	GDA	SLS	BLS	GDA	SLS	BLS
Revenue	89.534	89.571	89.571	78.964	79.311	83.142
Nb iter	69	34	30	527	32	24
Time (s)	0.636	0.376	1.656	14	1.3	41

The table shows that (BLS) gives the best solution. In the same time its complexity is high and grows significantly with the size of the problem. (SLS) presents an interesting trade-off between the quality of the solution and the computation time. The MPEC introduced in this paper enables the study of the effect of services' configuration (number of services, number of nests) and the nesting coefficients (correlation between services) on the services' tariffs policy. Future works will consider challenging problems like including discrete capacities design and multi-class setting where the services' disutility perception varies between customers.

References

[1] S. Anderson, A. de Palma, J. Thissei, *Discrete choice theory of product differentiation*, in MIT Press, 1992.

[2] S. Bekhor, L. Reznikova and T. Toledo, *Application of cross-nested Logit route choice model in stochastic user equilibrium traffic* in Transportation Research Record, no. 2003, pp. 41-49, 2007.

[3] M. Ben-Akiva, S. Lerman, *Discrete choice analysis: Theory and application to travel demand* Cambridge, MA: MIT Press, 1985.

[4] G.E Cantarella , A. Carteni , S. De Luca, *Stochastic equilibrium assignment with variable demand: Theoretical and implementation issues*, European journal of operational research, vol. 241, pp. 330-347, 2014.

[5] M. Chen, and A. Alfa, *Algorithms for solving Fisk's stochastic traffic assignment model*. Transportation Research B, vol. 25, 405-412, 1991.

[6] S. Erlander, N. Stewart, *The gravity model in transportation analysis: Theory and extensions* Utrecht: VSP, 1975.

[7] C. Fisk, *Some developments in equilibrium traffic assignment*, Transportation Research B, vol. 14, pp. 243-255, 1980.

[8] R. Hassin, M. Haviv, *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, in Kluwer, 2003.

[9] Y. Hayel and D. Quadri and T. Jimenez and L. Brotcorne, *Decentralized optimization of last-mile delivery services with non-cooperative bounded rational customers*, in Annals of Operations Research, on line first DOI 10.1007/s10479-014-1647-x, 2014.

[10] ZQ. Luo, JS. Pang, D. Ralph, *Mathematical programs with equilibrium constraints*, Cambridge University Press, 1996.

[11] D. MacFadden, *Quantal choice analysis: A survey*, in Annals of Economic and Social Measurement, vol. 5, pp. 363170, 1976.

[12] Q. Meng, D.H. Lee, H. Yang, H.J. Huang, *Transportation network optimization problems with stochastic user equilibrium constraints*, Transportation research record: Journal of the transportation research board, pp. 113-119, 2004.

[13] Y. Nesterov. *Gradient methods for minimizing composite objective function*, Core discussion papers, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

[14] Y. Sheffi, *Urban transportation networks*, Prentice-Hall, Inc., 1985.

[15] H. Yang, *Sensitivity analysis for queuing equilibrium network flow and its application to traffic control*, Mathematical and computer modeling, vol. 22, no. 4, pp. 247-258, 1995.

[16] K. Zhang, H.S. Mahmassani, C.C. Lu, *Dynamic pricing, heterogeneous users and perception error: Probit-based bi-criterion dynamic stochastic user equilibrium assignment*, Transportation Research Part C, vol 27, pp. 189-204, 2014.